Cape Breton University

KGC
The Knowledge Graph Conference

Creating a Healthcare Knowledge Graph From Statistical Open Data

HEALTHCARE AND LIFE SCIENCES SYMPOSIUM

Enayat Rajabi

enayat_rajabi@cbu.ca

May 2nd, 2022

Presenter

- Assistant Professor of Data Analytics, Cape Breton University, NS, Canada
- Adjunct Professor at the Faculty of Computer Science, Dalhousie University, Canada
- Canada NSERC grant holder

Open Statistical Data

- Publishing open statistical data is getting increased attention on the Web at various levels of governments.
- Open government data increases government transparency, and accountability, contributes to economic growth and improves administrative processes.
- A variety of open statistical data are published in various domains, including public health.
- The lack of meaning behind the statistical data makes it impossible to form a network and link this kind of data to infer, create and query knowledge.

A Knowledge Graph for Open Statistical Data

- Interconnectivity between isolated open datasets gives a machine much information to work with, thereby strengthening its ability to deduce relations and infer meaning.
- The design of a knowledge graph allows machines to draw new inferences that enrich the information available to users.
- Using classes and relationships, designing a knowledge graph links arbitrary entities or concepts from different domains.

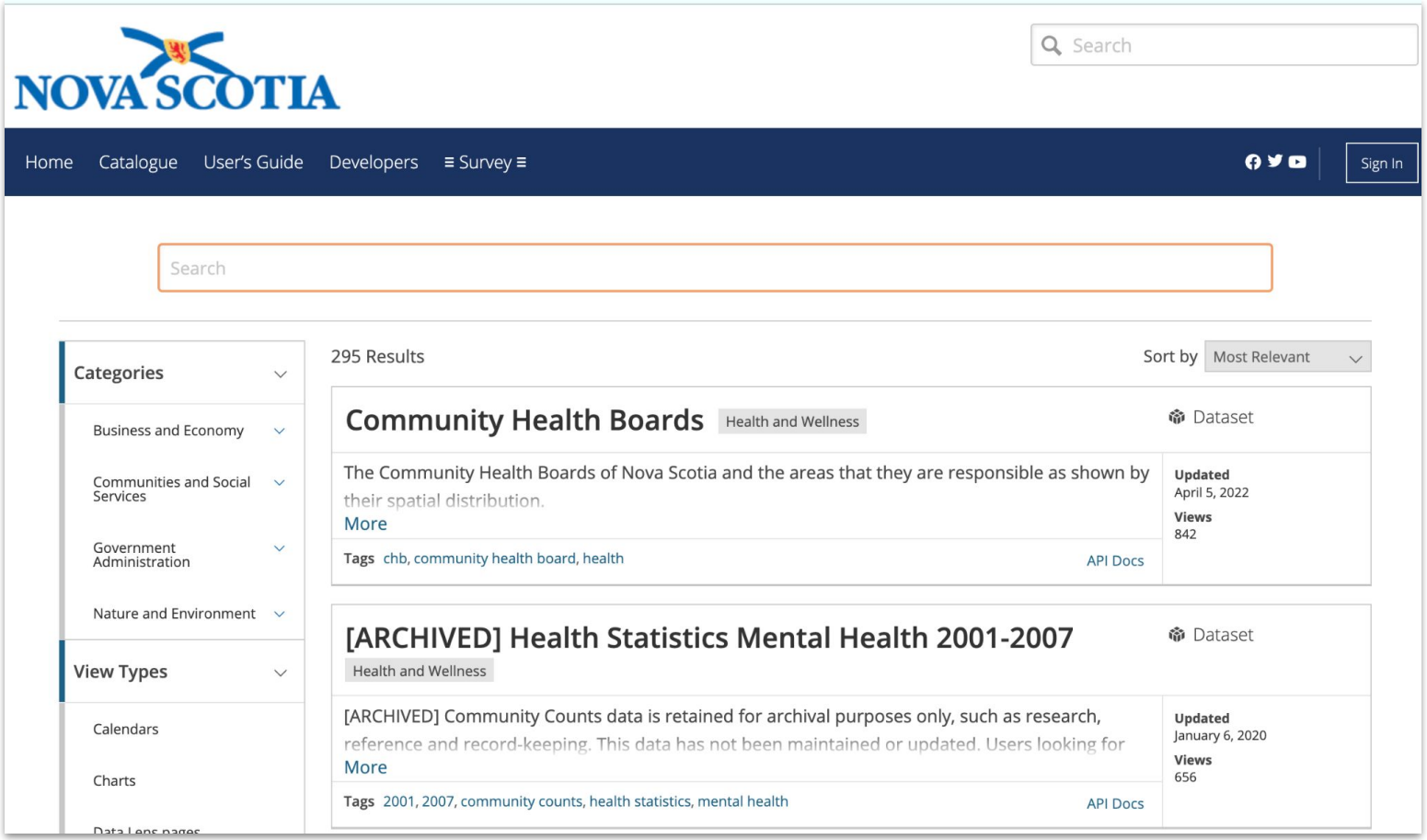
Canada Open Data

- Overall, Canada has 11 provinces and territories with approximately 11,771 published datasets in different domains ranging from “Business and Economy” to “Health and Wellness”.
- Each province publishes open statistical datasets via a provincial open data portal in different formats, including CSV, JSON, and Excel.
- A few of them allow users to export data in RDF format, they do not follow the Linked Data vocabulary five-star standards, and the open data portals are usually not linked to one another.

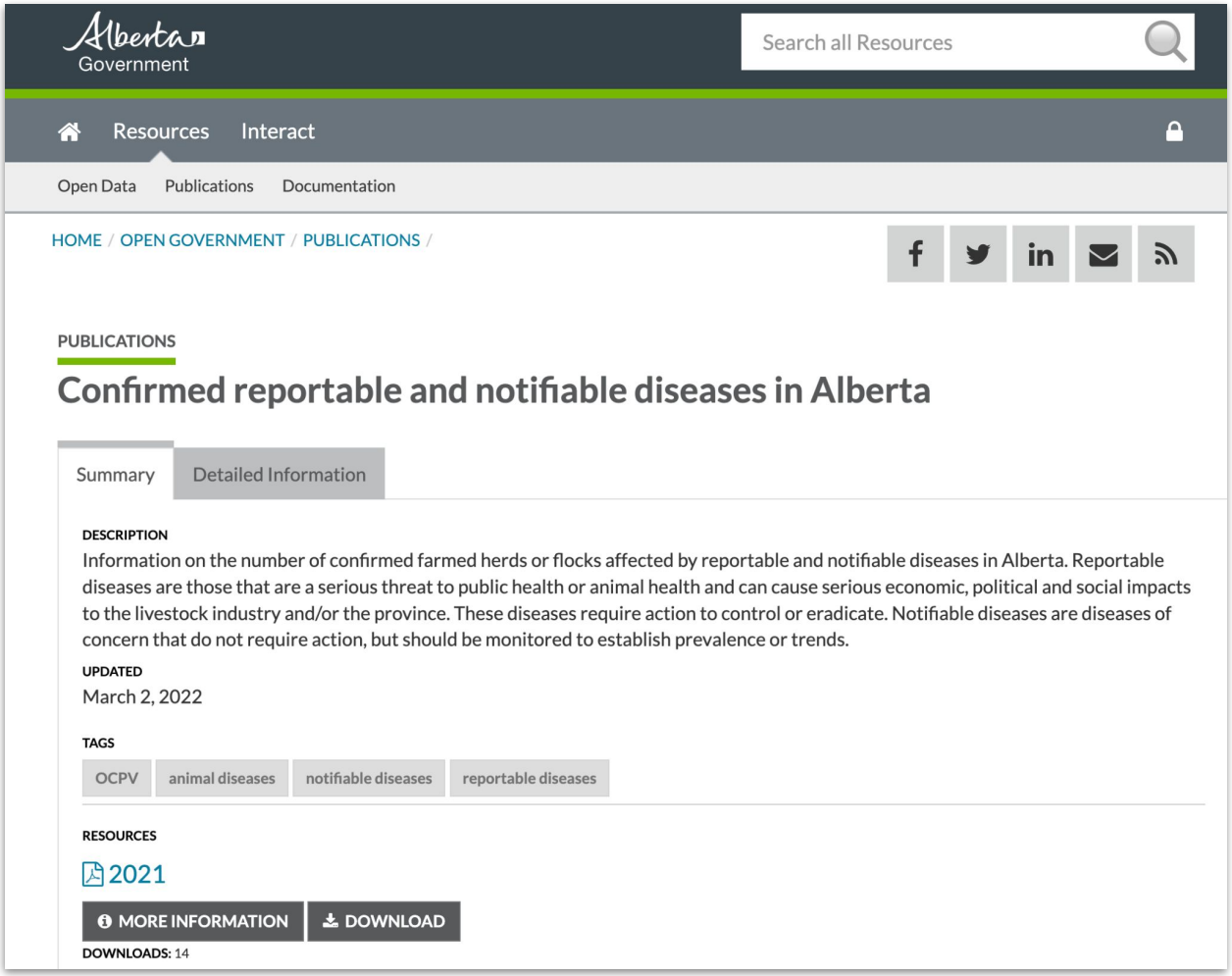
Canada Open Data (cont.)

- The datasets are isolated within or among the data portals, while many are conceptually linked.
- Statistical data regarding diseases in a province in different years should be manually analyzed to answer some important questions like: "Which viral diseases had the most number of cases in a province in 2017?".

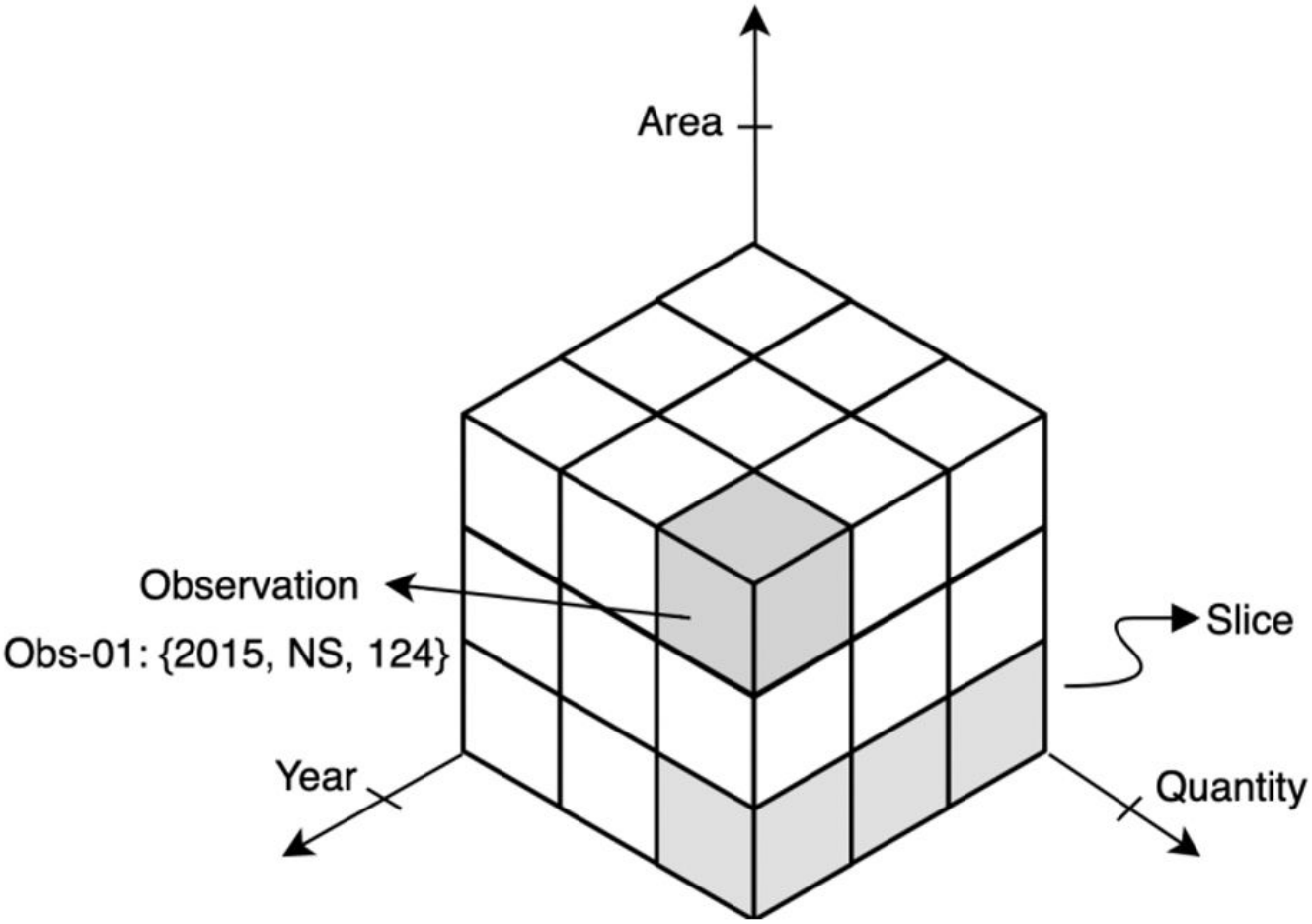
Provincial Open Data - Nova Scotia



Provincial Open Data - Alberta



Data Model for Statistical Data



Metadata in Open Statistical Data

Notifiable Diseases Counts and Rates 2005-2017

View Data | Visualize | Export | API | ...

Health And Wellness

Overall counts and rates (per 100,000 population) of notifiable diseases reported in Nova Scotia for 2005-2017.

Updated
December 5, 2018

About this Dataset

Mute Dataset

Updated
December 5, 2018

Data Last Updated
December 5, 2018

Metadata Last Updated
December 5, 2018

Date Created
November 17, 2015

Views
1,925

Downloads
172

Data Provided by
(none)

Dataset Owner
Open Data Nova Scotia

Detailed Metadata

Department	Health and Wellness
Geographic Region Name	Nova Scotia
Language	eng
Frequency	Annually
Time Period Coverage	2005-01-01 through 2017-12-31
Usage Considerations	Current as of November 19, 2018. 1. For 2005-2008, chronic and unspecified hepatitis B were reported together.; 2. For 2009-2016, Group A Streptococcal Disease Invasive are classified as 'Severe' or 'non-Severe'.; 3. Only diseases that have reported cases in the last 10 years are included.; 4. 2014: One case of Clostridium difficile and one case of MRSA did not report age group. 2015: 2 cases of MRSA and 1 case of Lyme Disease did not report age group. 2016: 1 case of MRSA and 2 cases of chlamydia did not report age group; 5. 2014: Two cases of chlamydia did not report sex. 2015: 3 cases of chlamydia did not report sex. 2016: 3 cases of chlamydia did not report sex.; 6. 2015: 5 HIV cases did not report zone. 2016: 2 HIV cases did not report zone.; 7. 2017: One case of MRSA did not report age: Two cases of Chlamydia did not report gender.
Related Documents	http://novascotia.ca/dhw/populationhealth/ ; http://novascotia.ca/dhw/populationhealth/diseases-and-conditions-A-Z.asp; http://novascotia.ca/dhw/cdpc/cdc/

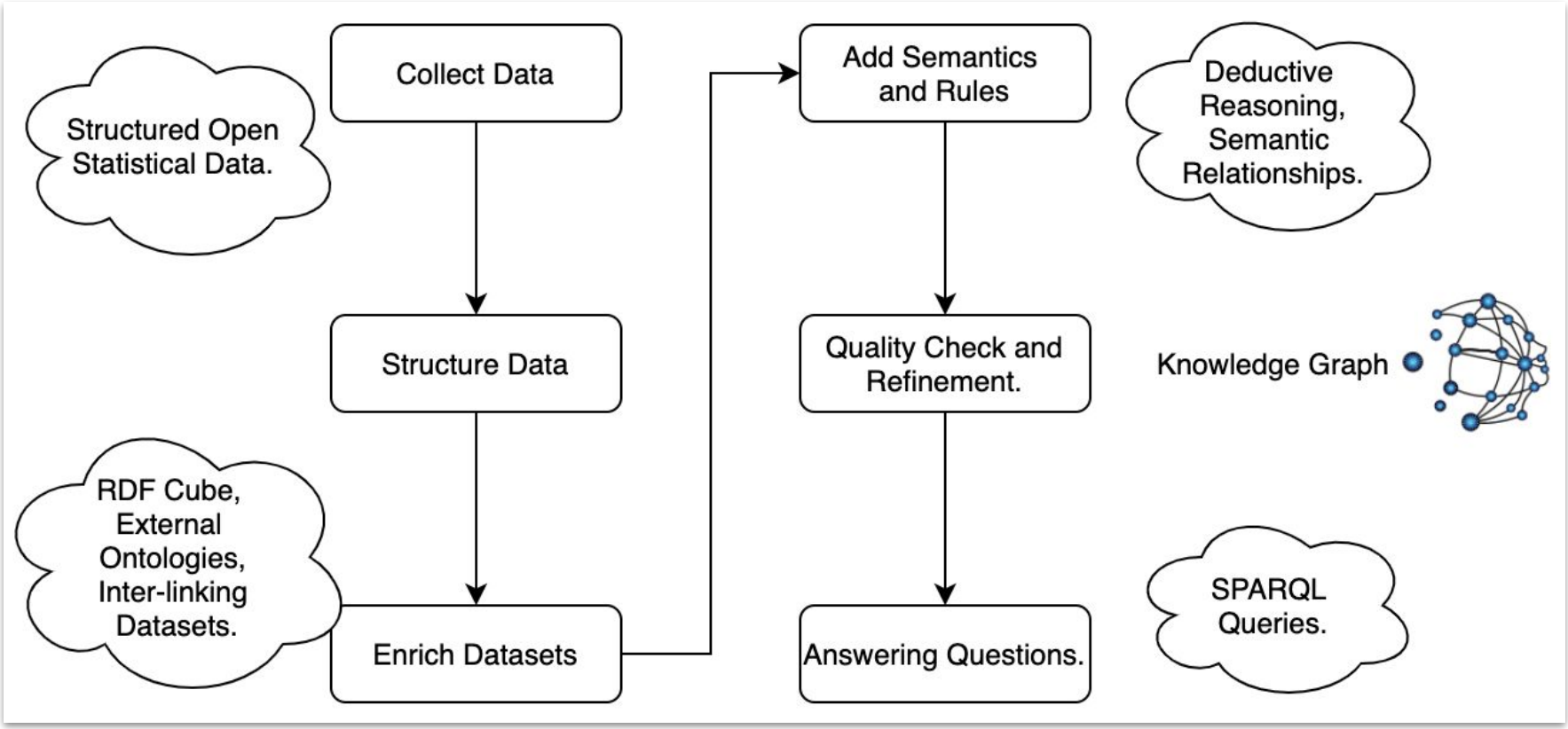
Open Statistical Data: Observations

Year	Disease	Number of Cases	Rate per 100,000 population
2005	Acquired Immune Deficiency Syndrome	5	0.5
2006	Acquired Immune Deficiency Syndrome	13	1.4
2007	Acquired Immune Deficiency Syndrome	5	0.5
2008	Acquired Immune Deficiency Syndrome	6	0.6
2009	Acquired Immune Deficiency Syndrome	2	0.2
2010	Acquired Immune Deficiency Syndrome	5	0.5
2011	Acquired Immune Deficiency Syndrome	4	0.4
2012	Acquired Immune Deficiency Syndrome	2	0.2
2013	Acquired Immune Deficiency Syndrome	0	0
2014	Acquired Immune Deficiency Syndrome	2	0.2
2005	Hepatitis B - Acute	10	1.1
2006	Hepatitis B - Acute	8	0.9
2007	Hepatitis B - Acute	9	1

< Previous Next >

Showing Rows 1 to 13 out of 689

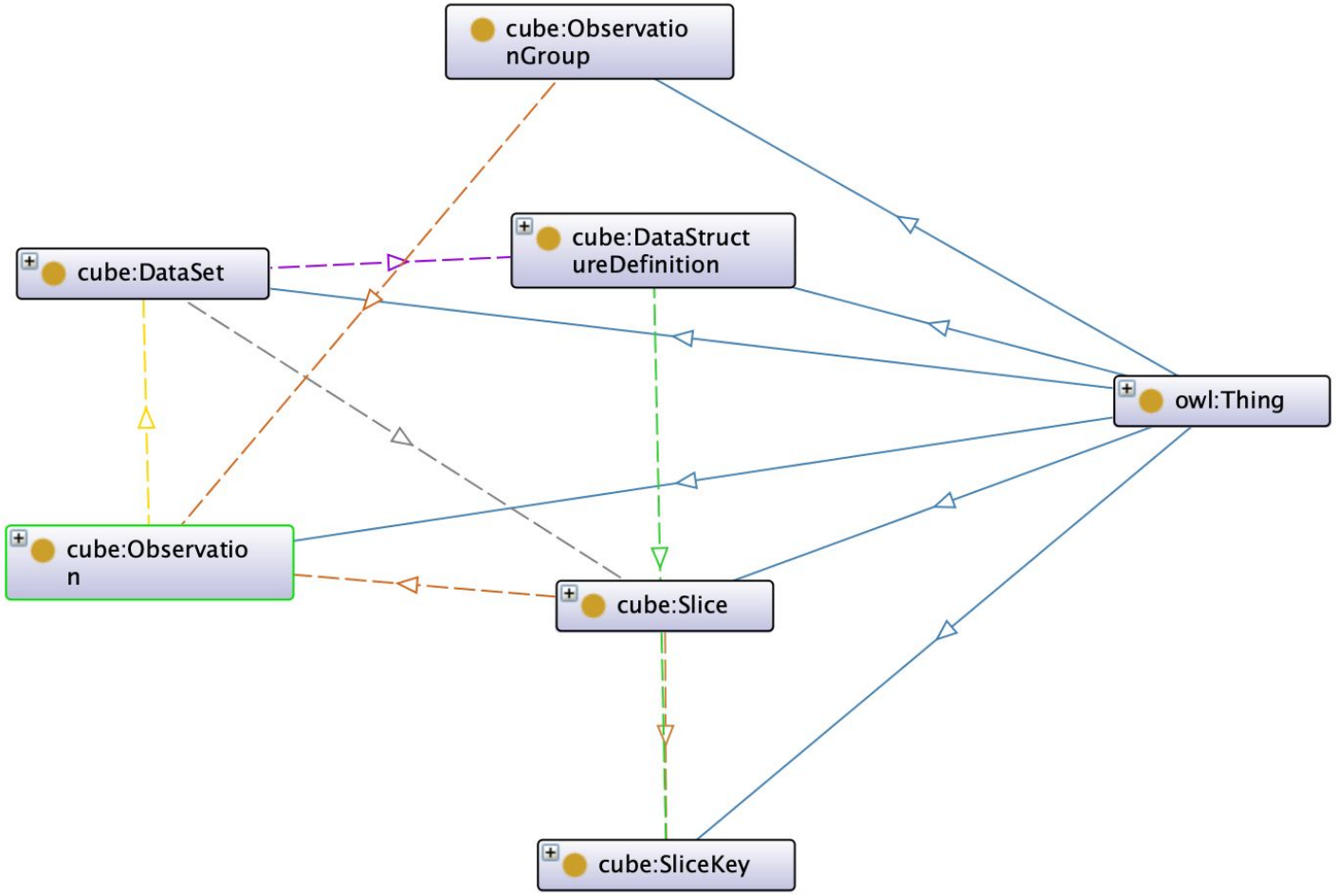
Knowledge Graph Construction Process



Creating Knowledge Graph for Disease Datasets

- Data collection: Used Python to download Nova Scotia datasets via Socrata API
- Data preparation: Analyzed the dataset and found 21 disease datasets
- Ontology: Created an ontology to struct the disease datasets
- Enrichment: Connected the datasets to the disease ontology (https://disease-ontology.org/)
- Semantic rules: Created SWRL rules in the knowledge graph to answer queries
- Published the constructed RDF knowledge graph on Zenodo website https://zenodo.org/record/5517236#.YmuvzvPML6Y

Ontology based on RDF Cube Library



Query Answering

- What are the viral infectious diseases in Nova Scotia?

```
SELECT ?disease_label ?disease_parent ?numofcases ?year
{
    ?observation eg:hasdisease ?disease.
    ?observation rdfs:label ?disease_label.
    ?disease rdfs:label ?disease_parent.
    FILTER regex(?disease_name,
                 "viral infectious disease", "i")
    ?observation eg:numberofcases ?numofcases.
    ?observation dimension:refPeriod ?year.
}
```

Conclusion and Next Steps

- Constructed a knowledge graph for open statistical data using the Semantic Web technologies and tools (RDF, SWRL, Protégé).
- Connected 21 statistical datasets in disease domain.
- The collected datasets had same structure. However, this is not the case in the other datasets.
- There are two challenges:
  - How to connect different datasets with different structures?
  - How to connect similar datasets in different provinces?